Using Perturbative Methods for Magnitude Tables in Statistical Disclosure Control

SCB

Sveriges officiella statistik

2024

Using Perturbative Methods for Magnitude Tables in Statistical Disclosure Control

Producer Statistics Sweden, Economic
Statistics and Analysis
SE-701 89 Örebro, Sweden
+46 10-479 40 00

Enquiries Nils Adriansson
+46 10-479 42 78
nils.adriansson@scb.se

Radka Sabolová
+46 10-479 61 37
radka.sabolova@scb.se

URN:NBN:SE:SCB-2024-UFFTBR2407_pdf

This publication is only available in electronic form on www.scb.se

Introduction

The data in the tables published for the Research and development (R&D) survey in the business enterprise sector and private non-profit sector are protected by using suppression of cells containing potentially sensitive data, i.e., cell suppression. Consequently, statistical information is withheld from users and their needs are not met. Especially tables presenting statistics by industry, or industry combined with other domains of interest such as size class, are supressed, diminishing the usefulness and relevance of the published statistics.

To avoid cell suppression, key respondents are asked to sign a waiver, giving consent to publishing cells containing their values. This means more re-contact with the organisation which needs to decide whether publishing the reported information can pose a risk to the organisation. These waivers are often left unanswered or, when answered, consent is often not given to disseminate cells where an organisation's data potentially can be disclosed.

Extensive work for disclosure limitation is performed to ensure no organisation is disclosed. Primary suppression of cells is done by identifying sensitive cell using the p% rule. This often results in further cells needing suppression, secondary suppression, to ensure that the primary suppressed cells cannot be backtracked. Additionally, internationally published tables often use different or more detailed aggregates than the nationally published tables. This requires extensive manual effort and results in more cells being suppressed.

Against this background, the R&D surveys for the business enterprise sector and private non-profit sector are the first surveys by Statistics Sweden where a perturbative method is used as a disclosure limitation technique for magnitude tables.

Description of the method

The so-called EZS-method is a simple perturbative method for magnitude tables that was introduced in Evans, Zayatz and Slanta (1998). This method adds noise to microdata and thus ensures additivity of tables and preserves links among tables.

Each object (organisation) in the sample is assigned two random values: direction of perturbation (+1 or -1) and noise factor (in percentages), generated from some chosen distribution. Note that these quantities remain confidential. The perturbed values are then computed as follows:

$$perturbed\ value = original\ value * (1 + direction * noise\ factor/100),$$

where both the direction and noise factor are applied to all values reported by the object. The distribution of directions of perturbation is chosen so that it is symmetric around 0 and thus does not introduce any consistent bias.

Cells containing only one object or cells with one dominant contribution are expected to obtain larger amount of noise as we aim to protect the individual objects in these cells. Noise in cells with many smaller contributions tend to cancel out and perturbed cell values are expected to be closer to their original, unperturbed counterparts.

To reduce the overall amount of noise added to the data we implemented the balancing procedure proposed in Massell and Funk (2007). This method is applied only to cells that do not have any disclosure risk according to a used disclosure rule and is not compromising level of protection.

To this end a single table and balancing sub-table (in our case single variable) that would steer the procedure need to be chosen. During the balancing procedure the random direction for specific objects can be changed to limit aggregated noise, and the new direction then applies to all values for that specific object. The balancing procedure is run at the most detailed level, e.g., the most detailed combination of domains.

The method uses the ‘greedy algorithm’ to minimise the total noise in safe cells. The values in safe cells are ordered from the largest to the smallest. The largest observation in a cell keeps its randomly assigned direction. The other objects’ directions might be altered so that the running noise total at each step is minimised, and the cumulative sum in the cell is therefore closer to the original unperturbed cumulative sum. This is achieved by selecting the direction to be the opposite of the sign of the running total noise in the cell. In unsafe cells (primary suppressions) the unbalanced version of the method is being used. Cells that would be considered secondary suppression are not considered being sensitive and thus are object to balancing. Note that the balancing procedure is carried out in one of the tables, but its results affect all tables.

Example

This short example illustrates the balancing procedure for the EZS-method. Let us look at one safe cell where five different objects contribute. All the objects were randomly assigned directions and random noise. Unperturbed cell total is equal to 2 000 while an unbalanced perturbed total for this cell equals 2 068.08, which is an increase by 3.4 percent. As this is a safe cell according to the used disclosure rule, we would like to alter the noise direction for some objects so that an estimate with lower amount of noise could be published. The two largest objects keep their randomly assigned directions which are opposite. As follows from the description above this set of directions results in a less perturbed cumulative sum in the cell. However, the third largest object F3 changes its direction so that the balanced perturbed cumulative sum is closer to the unperturbed sum. The two remaining objects, F4 and F5, keep their original directions. The balanced cell total is now 1 990.92 which amounts to a decrease of 0.45 percent in relative noise.

Table 1. Illustration of the EZS method, notice the change of direction for object F3.

Firm	Noise	Direction (original)	Value	Cumulative sum	Unbalanced	Direction +1	Direction -1	New direction
F1	10.94	1	1 000	1 000	1109.4			
F2	13.77	-1	450	1 450	1497.44	1 621.37	1 497.44	-1
F3	12.86	1	300	1 750	1 836.02	1 836.02	1 758.86	**-1**
F4	11.63	-1	200	1 950	2 012.76	1 982.12	1 935.60	-1
F5	10.65	1	50	2 000	2 068.08	1 990.92	1 980.27	1

Results and practical application

The method was implemented and tested on 2021 data from the survey R&D in the business enterprise sector. Noise factors were generated from a chosen distribution with values between *a* and *b*. Different choices of tables and variables were considered in the balancing procedure: variables with the highest number of supressed cells as well variables representing a significant part of the published total intramural R&D expenditure. The results varied, as the number of safe and unsafe cells fluctuated depending on the observed values, demonstrating that the outcome of the balancing procedure is highly sensitive to these factors.

Statistics Sweden investigated the effect of the EZS-method on one of the main variables, total intramural R&D expenditure. The results varied depending on the choice of method (balanced or unbalanced), variable used for balancing, and rounding. In order to evaluate the method, the distribution of the relative differences (in absolute value) between the original unperturbed estimates and the estimates based on the perturbed values was compared. This was done separately for unsafe and safe cells. We expect to see more relative noise in the unsafe cells as these needs to be sufficiently protected, whereas the safe cells are expected to exhibit smaller differences. Thus, to be able to publish the estimates for the safe cells as close as possible to their unperturbed values. The effect of perturbation on the total intramural R&D expenditures varied, for the unbalanced version of the EZS method increased the total by 4.8 percent, various choices of balancing variables resulted in an increase between 0.7 and 5.6 percent.

Histograms summarising these results for one chosen balancing variable are reported below. X-axis represents the percentage of noise added to cells (in absolute values), y-axis represents proportion of cells. The top panel shows that safe cells (including cells that would have been subject to secondary suppression) received smaller amounts of noise. Almost one quarter of these cells received almost no noise. Approximately 15 percent of the safe cells exhibited more substantial percentage changes as a result of the perturbation process. The histogram for unsafe cells (lower panel) shows a different distribution, with large number of cells receiving noise at least *a* percent, i.e., the minimum level of noise for each object. It is worth noting that, due to rounding to millions of Swedish kronor, the noise effects were diminished for some cells, as their rounded perturbed values remained the same as their rounded unperturbed values (as seen in the tallest bar around x=0 in the lower panel).

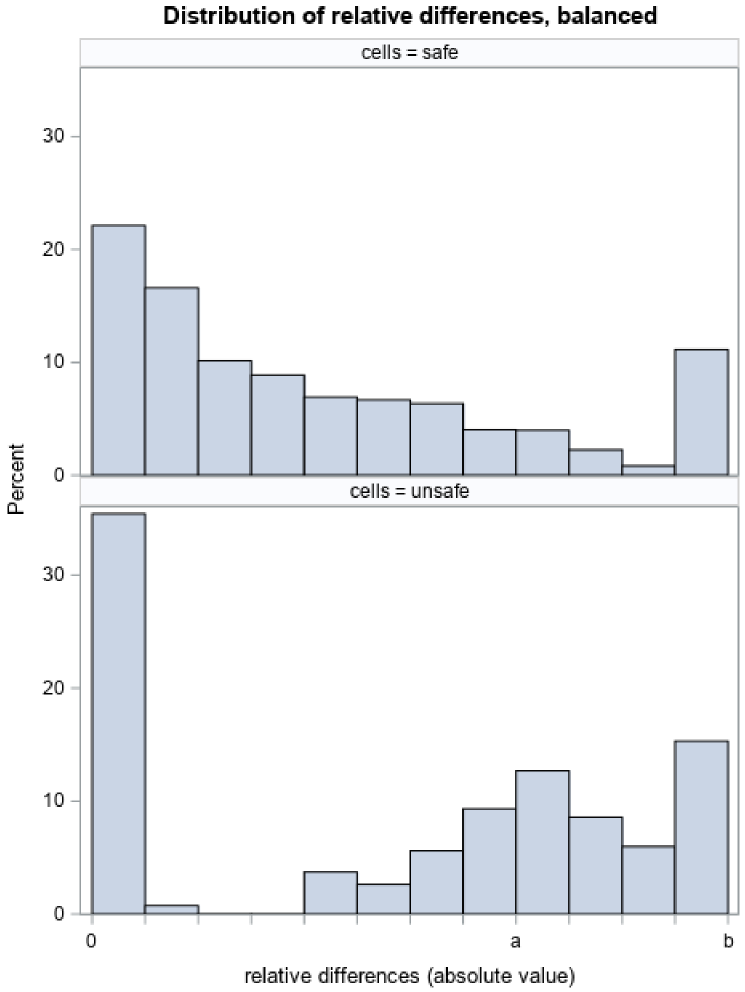

Figure 1. Distribution of relative differences, balanced.

Main findings

The method was tested using 2021 data from the survey R&D in the business enterprise sector and was successfully used for 2023 data published in autumn 2024. The method allows for disseminating tables with approximate values without supressing any cells in a consistent way and is easy to implement without need for any specialised software. As the noise is added to underlying microdata the method works well even for tables with high dimensions or tables with hierarchies and is much less time consuming than cell suppression.

The results from the above-described implementation on 2021 data does not contain any information on how noise affected data published for reference period 2023. The amount of noise applied to data might differ and relative differences depend on the observed values and other factors. Moreover, other changes have been implemented in the design of this survey which affects the published results.

References

T. Evans, L. Zayatz and J. Slanta, Using Noise for Disclosure Limitation of Establishment Tabular Data, Journal of Official Statistics, Vol 14, No. 4, 1998, pp. 537-551.

P. B. Massell and J. M. Funk, Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata, paper presented at the ICES-III, June 18-21, 2007, Montreal, Quebec, Canada.